Yash Akhauri

Research Scientist at Intel Labs
Hardware Software Co-Design, Mobile Computing,
Heterogeneous Architecture, Machine Learning

Cloud Systems Research
(+91) 7891512802
akhauri.yash@gmail.com
akhauriyash.github.io

Education

2016–2020 : **B.E. in Electronics & Instrumentation**, *Birla Institute of Science & Technology*, Rajasthan.
Thesis: Exposing Hardware Building Blocks to Machine Learning Frameworks [arXiv]

Research Portfolio

2021 Rethinking Zero-Shot Neural Architecture Scoring With Evolutionary Algorithms,
Yash Akhauri, J. Pablo Muñoz, Nilesh Jain, Ravi Iyer, **Under Review**.

2021 Enabling One-Shot NAS With Automatic Super-Network Generation,
J. Pablo Muñoz, Nikolay Lyalyushkin, **Yash Akhauri**, Anastasia Senina, Alexander Kozlov, Nilesh Jain, **Practical-DL AAAI'22**.

2021 A Genetic Programming Approach To Zero-Shot Neural Architecture Ranking,
Yash Akhauri, J. Pablo Muñoz, Nilesh Jain, Ravi Iyer, **AIPLANS NeurIPS'21**.

2020 LogicNets: co-designed neural networks and circuits for extreme-throughput applications,
Yaman Umuroglu*, **Yash Akhauri***, Nicholas James Fraser, Michaela Blott, **FPL'20**.
Stamatis Vassiliadis (best paper) Award & DATE Special Session Presentation

2020 High-throughput dnn inference with logicnets,
Yaman Umuroglu, **Yash Akhauri**, Nicholas James Fraser, Michaela Blott, **FCCM'20**.

2019 Hadanets: Flexible quantization strategies for neural networks,
Yash Akhauri, **UAVision CVPR'19 Oral**.

2021 RHNAS: Realizable Hardware and Neural Architecture Search,
Yash Akhauri*, Adithya Niranjan*, J Pablo Muñoz, Suvadeep Banerjee, Abhijit Davare, Pasquale Cocchini, Anton A Sorokin, Ravi Iyer, Nilesh Jain, Under Review at MLSys.

2021 BLOOMREC: Bloom Filter Based Memory Efficient Recommendation System,
Gopi Krishna Jha, Anthony Thomas, Nilesh Jain, **Yash Akhauri**, Ravi Iyer, Tajana Simunic Rosing, Under Review at MLSys.

2021 BootstrapNAS: Automated Generation Of Super-Networks From Pre-Trained Models For Neural Architecture Search,
J. Pablo Muñoz, Nikolay Lyalyushkin, Daniel Cummings, Anastasia Senina, Chaunté W Lacewell, **Yash Akhauri**, Alexander Kozlov, Nilesh Jain, Anthony Sarah, Under Review at MLSys.

Patent Applications [Approved for new filing]

2021 A System For Universal Hardware-Neural Network Architecture Search (Co-Design) .

2021 Apparatuses, Methods, And Systems For Instructions For Structured-Sparse Tile Matrix Fma.

2021 Efficient HW-SW Co-Design Using AutoML In OneAPI .

2021 Novel Method For Neural Network Compression And Decompression For Efficient Compute And Bw Utilization On Xeons For Improved AI Performance.

2021 Methods And Apparatus To Modify Pre-Trained Models To Apply Neural Architecture Search.

2021 Methods, Systems, Articles Of Manufacture And Apparatus To Optimize Resources In Edge Networks.

Talks

2019 **Intel Demo Booth at CVPR'19**, *Speaker* — *Long Beach, CA.*
Presented two demo talks on HadaNets

2019 **Intel AI DevCon**, *Poster - Oral* — *San Francisco, CA.*
Neural Network Quantization

2018 **Wolfram Technology Conference**, *Speaker* — *Champaign, IL.*
Introducing Hadamard Binary Neural Networks

2018 **Intel AI Meetup**, *Speaker* — *Delhi, IN.*
Scaling AI To Build An Intelligent World - Intel Case study

Experience

May'20 – present **Research Scientist**, *Cloud Systems Research - Intel Labs* — *Bangalore, India.*

- **Dynamic Inference Optimization**: Working on a closed-loop framework to dynamically optimize cache, memory bandwidth and core allocation. Utilizing dynamic depth classifiers to maximize inference performance at minimal model switching cost.
- **DLRM Optimization**: Formulated a NAS strategy for Deep Learning Recommendation Models (DLRM) and studying static cache and memory bandwidth allocation along with model switching with pareto optimal DLRM models.
- **Zero Shot NAS**: Proposed a framework to represent Neural Architecture Ranking algorithms as genetic programs. Utilized evolutionary search on genetic programs to discover SoTA Zero Shot Neural Architecture Ranking programs.
- **Sparse Acceleration**: Enabled pruning of image classification and transformer models and its software acceleration on next generation Intel Xeon CPUs.
- **AutoML**: Enabled efficient and realizable co-design of configurable hardware accelerators with arbitrary neural network search spaces.
- **Neural Network Compression**: Proposed a clustering based non-uniform neural network weight quantization scheme to maximize accuracy and amortize memory bandwidth requirement on CPUs and a LUT based multi-stage decompression framework for FPGAs.

Aug'19 – May'20 **Visiting Scholar**, *Xilinx Research* — *Dublin, Ireland.*
Advisor: **Dr. Yaman Umuroglu**, *Senior Research Scientist, Xilinx Labs*

- **LogicNets**: Developed the LogicNets library, explored extremely sparse and quantized MLPs and convolutional networks with PyTorch. Developed a Verilog code generator to convert MLPs in PyTorch to Verilog netlist for FPGA synthesis.
- **FPGA4HEP-Brevitas**: Developed a library to demonstrate Brevitas quantization library on the Jet Classification and Regression task to CERN.

Jan'19 – Jul'19 **Research Intern**, *Uraniom* — *France (Remote).*

- **Semantic Segmentation**: Utilized DeepLabV3 to enable semantic segmentation of facial features.

Jun'18 – Jul'18 **Undergraduate Researcher**, *Wolfram Summer School* — *Massachusetts.*
Advisor: **Dr. Sebastian Bodenstein**, *Research Engineer, DeepMind*

- **Neural Network Quantization**: Conducted research on neural network quantization and implemented a custom quantized neural network library from scratch in the Wolfram Language.

Fellowships & Awards

Intel Labs Invention Disclosure Award	
Exceptional Reviewer Award	AIPLANS@NeurIPS
Stamatis Vassiliadis Award (Best Paper Award)	FPL'20
Intel Nervana Early Innovators Grant	$5000
Intel CVPR Travel Grant	$3000
Wolfram Student Aid (Full Scholarship)	$2400
KVPY Fellowship by Dept of Science and Tech., Govt. of India	